Compositional analysis of high throughput sequencing data

Greg Gloor

26 January, 2018

Contents

Chapter 1

About this document

1.1 Reproducing the analysis

1.2 R packages required

Chapter 2

Introduction

2.1 Outline of the material

Chapter 3

The nature of sequencing data

3.1 DNA sequencing describes a random sample of the environment

3.1.1 Instrument capacity

3.1.2 Example calculation of fragment number

3.2 DNA sequencing is not counting

3.3 Random processes in sequencing

3.4 Sequencing post processing

3.4.1 Mapping

3.4.2 OTU generation

Chapter 4

DNA sequencing data are compositions

4.1 High throughput sequencing generates compositional data

In the Chapter 3 we saw that the capacity of the sequencing instrument imposed an upper bound on the total number of fragments that could be obtained from a given sequencing run. We also saw that the process of sequencing is essentially a random sampling of an environment where the environment contains more fragments than can possibly be sequenced. Finally, the data obtained are read counts per genetic interval (gene or OTU) per sample.

The read counts per sample range from 0 to, as a maximum, the total number of reads in the sample. Thus the data are positive integer data with an arbitrary maximum. While the data have an arbitrary maximum, the majority of current tools assume the data are counts and ignore the arbitrary maximum constraint. This assumption is the basis of methods grounded in distributions such as the zero inflated Gaussian (ZIG) (Paulson et al. 2013), negative binomial (Robinson, McCarthy, and Smyth 2010) and Poisson based models (Auer and Doerge 2011). Recent benchmarking has demonstrated that such methods are unpredictable when dealing with highly sparse data (Thorsen et al. 2016) and do not control the false discovery rate (G. B. Gloor, Macklaim, et al. 2016; Hawinkel et al. 2017).

Data of this type are called count compositions, and a number of groups have started to work on developing appropriate methods to deal with high throughput datasets as count compostions (Friedman and Alm 2012; Fernandes et al. 2013, 2014; Lovell et al. 2015; Mandal et al. 2015; Kurtz et al. 2015; G. B. Gloor and Reid 2016; Erb and Notredame 2016; G. B. Gloor, Wu, et al. 2016; Tsilimigras and Fodor 2016; Washburne et al. 2017; T. Quinn et al. 2017; Silverman et al. 2017; T. P. Quinn et al. 2017; Kaul et al. 2017; Erb et al. 2017; Egozcue, Pawlowsky-Glahn, and Gloor 2018).

So what is compositional data, and what are its properties with respect to high throughput sequencing that make this an important issue?

4.2 Compositional data

Data from high throughput sequencing have the following properties; the data are counts, the data are non-negative, and the data has an upper bound imposed by the instrument because there is a limit to the number of fragments (and hence gene or OTU counts) that can be observed. This fits with the definition of compositional data: the data contains D features (OTUs, genes, etc), where the count of each feature is non-negative, and the sum of the parts is known (Aitchison 1986, pg25). Note that the data do not have to sum to a predetermined amount, it is sufficient that the sum of the parts be known and not be able to be exceeded.

A vector containing D features where the sum is 1 can be formally stated as: $\vec{X} = \{(x_1, x_2, x_3, \ldots x_D); x_i \geq 0; \sum_{x=1}^{D} = 1\}$. The sum of the parts is usually set to 1 or 100, but can take any value; i.e., any composition can be scaled to any arbitrary sum such as a ppm. The property of scaling to any arbitrary value is named an equivalence class and compositional data are equivalence classes (Barceló-Vidal, Martín-Fernández, and Pawlowsky-Glahn 2001). In the lexicon of high throughput sequencing the vector is the sample and the features are the OTUs or

genes or genomic intervals. The total sum is the total number of fragments observed for the sample.

Compositional data have a number of built-in pathologies: a negative correlation bias, sub-compositional incoherence, and spurious correlations. A proper analysis of compositional data must as a minimum account for these pathologies.

More formally, compositional datasets have the property that they are described by $D-1$ features (Aitchison 1986). In other words, if we know that all features sum to 1, then the value of any individual feature can be known by subtracting the sum of all other parts from 1, i.e., $x_D = 1 - \sum_{x=1}^{D-1}$.

Graphically, this means that compositional data inhabit a space called a Simplex that contains 1 fewer dimensions than the number of features. The distances between parts on the Simplex are not linear. This is important because all parametric statistical tests assume that differences between parts are linear (or additive). Thus, while standard tests will produce output, the output will be misleading because distances on the simplex are non-linear and bounded (Martín-Fernández et al. 1998). Chapter 5 on Data Transformations contains an intuitive demonstration of how data are moved to the Simplex when a the data are compositional.

It is not always apparent when the data are compositional. This is especially true for large multivariate datasets such as those generated in high throughput sequencing. Aitchison (1986) indicated that a compositionally appropriate analysis should fulfil a number of properties, and when these properties are not met with traditional analyses, the data is likely compositional.

- A compositionally appropriate analysis should be scale invariant, that is, the results should not depend on the total count or scale of the sample. There is substantial resistance to the idea the high throughput sequencing data are compositional, and indeed many analysts believe that the data can be made non-compositional with the 'correct' transform that restores the scale. The is belief is exposed to be false in Chapter 5.
- A compositionally appropriate analysis should also not depend on the order of the features in the dataset. This almost goes without saying, but is included because of the way that one particular transformation, the alr, was formulated.
- A compositionally appropriate analysis should exhibit subcompositional coherence, or the results of analysis of a sub-composition should be the same as for the entire composition. In practice, this is difficult to achieve, and we settle for least sub-compositional dominance where the distances between features in the full composition is equal to or greater than the distances in the sub-composition. In later chapters where we examine real datasets, I show how to determine if sub-compositional coherence and dominance are fulfilled by the analysis.

4.2.1 Negative correlation bias in compositions

The values of the parts of compositional datasets are constrained because of the constant sum, and this constraint has been known for a very long time. The features in a composition have a negative correlation bias since an increase in the value of one part must be offset by a decrease in value of one or more other parts. In the illustration above, we see that 'V' and 'M' are perfectly balanced on the fulcrum because they have the same mass. If M becomes heavier, then V will rise even though the mass of V has not changed. The same principle operates in compositional data. If V is the amount of money spend on vegetables, and M is the amount of money spent on meat, and the total food budget is a constant, then the only way that more meat would be consumed would be to spend less on vegetables. Therefore, the amount of money spent on V and M will be perfectly negatively correlated if the total food budget is constrained. This example generalizes to any number of items in the shopping basket as long as the total budget is constrained. When there are more items, then an increase in one item (say shoes) must be offset by a decrease in another item, but it could be a decrease in meat, vegetables or both.

4.2.2 Spurious correlations:

In addition to a negative correlation bias, compositional data has the problem of spurious correlation (Pearson 1897); in fact spurious correlation was the first troubling issue identified with compositional data. This phenomenon is best illustrated with the

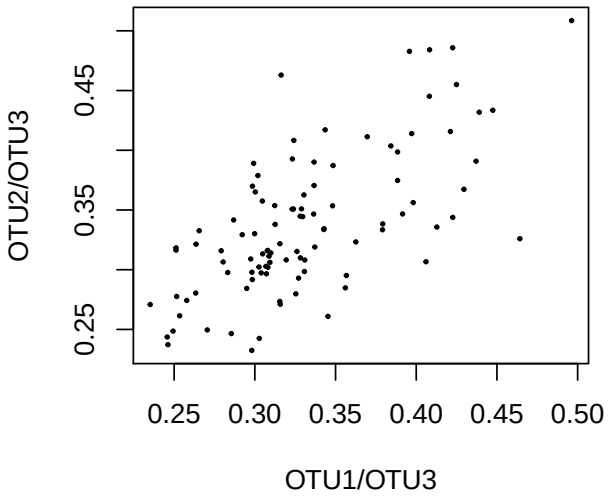

Figure 4.1: Spurious correlation in compositional data. Two random vectors drawn from a Normal distribution, were divided by a third vector also drawn at random from a Normal distribution. The two vectors have nothing in common, they should exhibit no correlation, and yet they exhibit a correlation coefficent of > 0.65 when divided by the third vector. See the introductory section of the Supplementary Information of Lovell (2015) for a more complete description of this phenomenon.

following example from Lovell et. al (2015), where they show how simply dividing two sets of random numbers (say abundances of OTU1 and OTU2), by a third set of random numbers (say abundances of OTU3) results in a strong correlation. Note that this phenomenon depends only on there being a common denominator.

```
n.obs <- 100
OTU.df <- data.frame(
    OTU1=rnorm(n.obs, mean=10, sd=1),
    OTU2=rnorm(n.obs, mean=10, sd=1),
    OTU3=rnorm(n.obs, mean=30, sd=4))
OTU.df <- transform(OTU.df,
    OTU1.over.OTU3= OTU1/OTU3,
    OTU2.over.OTU3= OTU2/OTU3)
plot(OTU.df$OTU1.over.OTU3,
    OTU.df$OTU2.over.OTU3, pch=19,
    cex=0.3,xlab="OTU1/OTU3",
    ylab="OTU2/OTU3")
```

4.2.3 Sub-compositions

Compositional data have the third property of sub-compositional incoherence of correlation metrics as illustrated in Chapter 5. That is, *correlations calculated on compositional datasets are unique to the particular dataset chosen* (Aitchison 1986). This is problematic because high throughput sequencing experimental designs are *always* sub-compositions. Inspection of papers in the literature provide many examples. For example, in the 16S rRNA gene sequencing literature it is common practice to discard rare OTU species prior to analysis and to re-normalize by dividing the counts for the remaining OTUs by the new sample sum. It is also common to use only one or a few taxonomic groupings to determine differences between experimental conditions. In the case of RNA-seq only the fraction of RNA of interest is sequenced, usually mRNA but other sub-fractions such as miRNA may be sequenced. All of these practices expose the investigator to the problem of non-coherence between sub-compositions. We must use compositionally-appropriate measures of correlation—more formally, we are attempting to find features that are compositionally associated. Compositional association as a more restricted measure of correlation and is explained more completely in the chapter on data transformations.

To summarize, compositional data has the following pathologies:

- The negative correlation bias means that any negative correlation observed in compositional data must be treated as suspect because it could arise simply because a different feature (or features) changed their abundance. There is currently no theoretically valid approach to identify true negative correlations in compositional data (Egozcue, Pawlowsky-Glahn, and Gloor 2018).
- The spurious correlation problem means that we can observe apparent postive correlations simply by chance. I describe recent work that shows that spurious correlation is tractable.
- The sub-compositional incoherence of correlation is perhaps the most insidious property, but also the easiest to recognize. Here the correlation depends on the *exact* set of features present in the dataset. If the observed correlations change when the data are subset, then sub-compositional incoherence is in play.

Thus, one major reason to use compositional data methods is that you are more likely to report robust results, and the later practical chapters demonstrate the robustness of a compositional data analysis.

Practically speaking the negative correlation bias, the occurrence of spurious correlation, and the prob-

lem of sub-compositional incoherence means that *every microbial correlation network that has ever been published is suspect*, as is *every gene co-occurrence or co-expression network* unless compositionally appropriate compositional association metric was used (Lovell et al. 2015; Erb and Notredame 2016; T. P. Quinn et al. 2017). These approaches themselves have limitations and as originally constituted cannot deal with sparse data. However, recasting the data from count compositions to probability distributions allows these methods to be adapted to sparse data with some success (Bian et al., n.d.; T. P. Quinn et al. 2017).

4.3 So can I analyze compositional data? How?

Much of the high throughput sequencing analysis literature seems to assume that data derived from high throughput sequencing are in some way unique, and that purpose-built tools must be used. However, there is nothing special about high-throughput sequencing data from the point of view of the analysis. Fortunately, the analysis of compositional datasets has a well-developed methodology (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015; Van den Boogaart and Tolosana-Delgado 2013), much of which was worked out in the geological sciences.

Atichison (1986), Pawlsky-Glahn (2006), and Egozcue (2005), have done much work to develop rigorous approaches to analyze compositional data (Pawlowsky-Glahn and Buccianti 2011). The essential step is to reduce the data to ratios between the D. This step does not move the data from the Simplex but does transform the data on the Simplex such that the distances between the ratios of the features are linear. The investigator must keep in mind that the distances are between ratios between features, not between counts of features (re-read this several times to wrap your head around it). Several transformations are in common use, but the one I believe is most applicable to HTS data is the centred log-ratio transformation or clr, where the data in each sample is transformed by taking the logarithm of the the ratio between the count value for each part and the geometric mean count: i.e., for D features in sample vector $\vec{\mathbf{X}} = [x_1, x_2, x_3, \dots x_D]$:

$$\vec{\mathbf{X}}_{clr} = [log(\frac{x_1}{gX}), log(\frac{x_2}{gX}) \dots log(\frac{x_D}{gX})] \qquad (4.1)$$

where gX is the geometric mean of the features in sample $\vec{\mathbf{X}}$. The clr transformation is formally equivalent to a matrix of all possible pairwise ratios, but is a more tractable form. The clr transformation is not perfect by any means, but when there are large numbers of features the properties of the clr approach the ideal isometric log-ratio transformation or ilr. In the context of high throughput sequencing, where there are often hundreds or thousands of features the clr and the ilr have nearly indistinguishable properties.

The properties of the clr transformation are demonstrated in the Chapter **??**.

Chapter 5

Data transforms in high throughput sequencing

Chapter 6

Exploring compositional data: the compositional biplot

Chapter 7

References

Aitchison, J. 1986. *The Statistical Analysis of Compositional Data*. London, England: Chapman & Hall.

Auer, Paul L, and Rebecca W Doerge. 2011. "A Two-Stage Poisson Model for Testing RNA-seq Data." *Statistical Applications in Genetics and Molecular Biology* 10 (1).

Barceló-Vidal, Carles, Josep A Martín-Fernández, and Vera Pawlowsky-Glahn. 2001. "Mathematical Foundations of Compositional Data Analysis." In *Proceedings of IAMG*, 1:1–20.

Bian, Gaorui, Gregory B Gloor, Aihua Gong, Changsheng Jia, Wei Zhang, Jun Hu, Hong Zhang, et al. n.d. "The Gut Microbiota of Healthy Aged Chinese Is Similar to That of the Healthy Young." *mSphere* 2 (5):e00327–17. https://doi.org/10.1128/mSphere.00327-17.

Egozcue, JJ, and V. Pawlowsky-Glahn. 2005. "Groups of Parts and Their Balances in Compositional Data Analysis." *Mathematical Geology* 37 (7). Springer:795–828.

Egozcue, Juan José, Vera Pawlowsky-Glahn, and Gregory B. Gloor. 2018. "Linear Association in Compositional Data Analysis." *Austrian Journal of Statistics* in press.

Erb, Ionas, and Cédric Notredame. 2016. "How Should We Measure Proportionality on Relative Gene Expression Data?" *Theory in Biosciences* 135 (1):21–36.

Erb, Ionas, Thomas Quinn, David Lovell, and Cedric Notredame. 2017. "Differential Proportionality - a Normalization-Free Approach to Differential Gene Expression." *bioRxiv*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/134536.

Fernandes, Andrew D, Jean M Macklaim, Thomas G Linn, Gregor Reid, and Gregory B Gloor. 2013. "ANOVA-Like Differential Expression (Aldex) Analysis for Mixed Population Rna-Seq." *PLoS One* 8 (7):e67019. https://doi.org/10.1371/journal.pone.0067019.

Fernandes, Andrew D, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. 2014. "Unifying the Analysis of High-Throughput Sequencing Datasets: Characterizing RNA-Seq, 16S RRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis." *Microbiome* 2:15.1–15.13. https://doi.org/10.1186/2049-2618-2-15.

Friedman, Jonathan, and Eric J Alm. 2012. "Inferring Correlation Networks from Genomic Survey Data." *PLoS Comput Biol* 8 (9):e1002687. https://doi.org/10.1371/journal.pcbi.1002687.

Gloor, Gregory B, Jean M Macklaim, Michael Vu, and Andrew D Fernandes. 2016. "Compositional Uncertainty Should Not Be Ignored in High-Throughput Sequencing Data Analysis." *Austrian Journal of Statistics* 45:73–87. https://doi.org/doi:10.17713/ajs.v45i4.122.

Gloor, Gregory B, and Gregor Reid. 2016. "Compositional Analysis: A Valid Approach to Analyze Microbiome High-Throughput Sequencing Data." *Can J Microbiol* 62 (8):692–703. https://doi.org/10.1139/cjm-2015-0821.

Gloor, Gregory B, Jia Rong Wu, Vera Pawlowsky-Glahn, and Juan José Egozcue. 2016. "It's All Relative: Analyzing Microbiome Data as Compositions." *Ann Epidemiol* 26 (5):322–9. https://doi.org/10.1016/j.annepidem.2016.03.003.

Hawinkel, Stijn, Federico Mattiello, Luc Bijnens, and Olivier Thas. 2017. "A Broken Promise: Microbiome Differential Abundance Methods Do Not Control the False Discovery Rate." *Briefings in Bioinformatics*. Oxford University Press, bbx104.

Kaul, Abhishek, Siddhartha Mandal, Ori Davidov, and Shyamal D Peddada. 2017. "Analysis of Microbiome Data in the Presence of Excess Zeros."

Front Microbiol 8:2114. https://doi.org/10.3389/fmicb.2017.02114.

Kurtz, Zachary D, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. 2015. "Sparse and Compositionally Robust Inference of Microbial Ecological Networks." *PLoS Comput Biol* 11 (5):e1004226. https://doi.org/10.1371/journal.pcbi.1004226.

Lovell, David, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. 2015. "Proportionality: A Valid Alternative to Correlation for Relative Data." *PLoS Comput Biol* 11 (3):e1004075. https://doi.org/10.1371/journal.pcbi.1004075.

Mandal, Siddhartha, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. 2015. "Analysis of Composition of Microbiomes: A Novel Method for Studying Microbial Composition." *Microb Ecol Health Dis* 26:27663.

Martín-Fernández, JA, C Barceló-Vidal, V Pawlowsky-Glahn, A Buccianti, G Nardi, and R Potenza. 1998. "Measures of Difference for Compositional Data and Hierarchical Clustering Methods." In *Proceedings of IAMG*, 98:526–31.

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. "Differential Abundance Analysis for Microbial Marker-Gene Surveys." *Nat Methods* 10 (12):1200–1202. https://doi.org/10.1038/nmeth.2658.

Pawlowsky-Glahn, V., and J. J. Egozcue. 2006. "Compositional Data and Their Analysis: An Introduction." *Geological Society, London, Special Publications* 264 (1):1–10. https://doi.org/10.1144/GSL.SP.2006.264.01.01.

Pawlowsky-Glahn, Vera, and Antonella Buccianti. 2011. *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

Pawlowsky-Glahn, Vera, Juan José Egozcue, and Raimon Tolosana-Delgado. 2015. *Modeling and Analysis of Compositional Data*. John Wiley & Sons.

Pearson, Karl. 1897. "Mathematical Contributions to the Theory of Evolution. – on a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs." *Proceedings of the Royal Society of London* 60:489–98.

Quinn, Thomas P., Ionas Erb, Mark F. Richardson, and Tamsyn M. Crowley. 2017. "Understanding Sequencing Data as Compositions: An Outlook and Review." *bioRxiv*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/206425.

Quinn, Thomas, Mark F Richardson, David Lovell, and Tamsyn Crowley. 2017. "Propr: An R-Package for Identifying Proportionally Abundant Features Using Compositional Data Analysis." *bioRxiv*. Cold Spring Harbor Labs Journals. https://doi.org/10.1101/104935.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1):139–40. https://doi.org/10.1093/bioinformatics/btp616.

Silverman, Justin D, Alex D Washburne, Sayan Mukherjee, and Lawrence A David. 2017. "A Phylogenetic Transform Enhances Analysis of Compositional Microbiota Data." *Elife* 6 (February):21887. https://doi.org/10.7554/eLife.21887.

Thorsen, Jonathan, Asker Brejnrod, Martin Mortensen, Morten A Rasmussen, Jakob Stokholm, Waleed Abu Al-Soud, Søren Sørensen, Hans Bisgaard, and Johannes Waage. 2016. "Large-Scale Benchmarking Reveals False Discoveries and Count Transformation Sensitivity in 16S RRNA Gene Amplicon Data Analysis Methods Used in Microbiome Studies." *Microbiome* 4 (1):62. https://doi.org/10.1186/s40168-016-0208-8.

Tsilimigras, Matthew C B, and Anthony A Fodor. 2016. "Compositional Data Analysis of the Microbiome: Fundamentals, Tools, and Challenges." *Ann Epidemiol* 26 (5):330–5. https://doi.org/10.1016/j.annepidem.2016.03.002.

Van den Boogaart, K Gerald, and Raimon Tolosana-Delgado. 2013. *Analyzing Compositional Data with R*. Springer, London, UK.

Washburne, Alex D, Justin D Silverman, Jonathan W Leff, Dominic J Bennett, John L Darcy, Sayan Mukherjee, Noah Fierer, and Lawrence A David. 2017. "Phylogenetic Factorization of Compositional Data Yields Lineage-Level Associations in Microbiome Datasets." *PeerJ* 5:e2969. https://doi.org/10.7717/peerj.2969.